JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS
VOLUME 5, ISSUE 02
MOPNTHLY JOURNALS
ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

THEORETICAL AND METHODOLOGICAL FOUNDATIONS FOR CREATING A LINGUISTIC BASE OF THE FERGANA VALLEY DIALECTS AND INTEGRATING THEM INTO THE NATIONAL CORPUS OF THE UZBEK LANGUAGE

Vosiljonov Azizbek Boxodirjon ugli
Teacher of Fergana State University
azizbekvosiljonov@gmail.com

Abstract

This article comprehensively analyzes the issues of systematic study of the linguistic characteristics of dialects distributed in the Fergana Valley, their digitization and integration into the national corpus of the Uzbek language. Within the framework of the research, phonetic, morphological and lexical units were identified based on field materials, and a multi-layered linguistic annotation model was developed. The processes of transcription, normalization, enrichment with metadata of dialectal texts and their adaptation to the corpus architecture were scientifically covered. The article shows the importance of determining areal differences, creating opportunities for linguostatistical analysis, and the dialectal module in applied language technologies. The results of the research suggest a methodological model for preserving and systematizing Fergana Valley dialects in a digital environment.

Keywords

dialectology, corpus linguistics, Fergana Valley dialects, linguistic base, annotation, transcription, areal feature, digitization, metadata.

Introduction. Territorial variants of the language are an integral part of the national language system. Dialects reflect the living layer of the history of the language, the ethnic and cultural experience of the people, and the stages of socio-cultural development. Therefore, their study is of great importance not only for dialectology, but also for general linguistics, ethnolinguistics, and sociolinguistics. In recent decades, corpus linguistics has strengthened the empirical basis of language research. Electronic corpora have made it possible to statistically and functionally analyze language units based on real speech materials. Work is also underway to create a national corpus in the Uzbek language. However, territorial speech variants, in particular, the dialects of the Fergana Valley, have not yet been fully and systematically included in the corpus. The Fergana Valley is distinguished by its historical and cultural layers, ethnic composition, and economic ties. The speech system in this region has internal areal differences and exhibits its own phonetic, morphological, and lexical characteristics. Therefore, the formation of Fergana dialects as a linguistic base and their integration into the national corpus is an urgent scientific task.

Linguistic description of the dialects of the Fergana Valley

Although the dialects of the Fergana Valley were formed within the framework of the Karluk dialect, their internal differential signs can be observed. In the phonetic system, the narrowing or widening of vowels, soft or hard pronunciation of some consonants, assimilation and dissimilation processes are observed. In some cases, variantization of phonemes is also observed. In the morphological system, variants of verb tenses and moods, phonetic adaptation of some suffixes, and parallel use of form variants are observed. The lexical layer is rich in terms specific to the regional lifestyle, ethnographic units, ancient and local words. These features differ areally, and in some regions the phonetic sign prevails, while in others the lexical

difference is more pronounced. Therefore, it is important to determine regional differentiation based on a linguogeographic approach.

Collection and systematization of dialectal material. The study was conducted on the basis of field materials. Speeches of representatives of different ages, professions and social strata were recorded. Oral speech materials were stored in audio format and later transferred to written form. A two-stage approach was used in the transcription process. At the first stage, phonetic transcription was carried out, preserving the sound characteristics of the dialect. At the second stage, normalization was carried out, creating a variant close to the literary language. This method provides a convenient opportunity for comparative analysis of dialectal units and statistical calculations.

After the texts were converted to digital format, linguistic tags were assigned to them. Each word was marked at the phonetic, morphological and lexical levels. In addition, the texts were enriched with metadata. The informant's age, gender, place of residence, date of recording and topic were indicated. This serves as an important source for sociological and linguostatistical analysis.

Linguistic annotation and corpus architecture

A multilayer annotation model was used to create the dialect module. During the annotation process, phonetic variants were separately identified, morphological forms were recorded based on grammatical labeling, and lexical units were separated by regional tags. The corpus architecture was developed in accordance with the general national corpus system. The dialect module is integrated as an independent layer, allowing the user to view literary language and dialect variants in parallel. The search system allows you to determine the frequency of a particular unit, show its territorial distribution, and analyze it based on context.

This approach allows you to conduct linguostatistical analysis, identify differentiation across regions, and study the functional load of dialect units.

Scientific and practical significance

Creating a linguistic database of the dialects of the Fergana Valley serves to preserve the territorial wealth of the Uzbek language and systematize it on a scientific basis. This database serves as a reliable empirical source for dialectological research. Also, dialectal data can be a necessary training resource for automatic speech recognition, automatic text analysis, machine translation and artificial intelligence systems. Models created taking into account the dialectal layer allow for more accurate processing of real speech. In the educational process, this linguistic base can be effectively used to teach regional speech features, conduct practical exercises with students, and conduct scientific research.

Conclusion: Creating a linguistic base of the Fergana Valley dialects and integrating them into the national corpus of the Uzbek language is one of the current directions of modern linguistics. The multilayer annotation model made it possible to systematically record dialectal units, analyze them statistically and functionally.

This approach will serve as a methodological basis for the future inclusion of other regional dialects in the corpus, the creation of a dialectal module on a republican scale, and the formation of a complete digital map of the Uzbek language. Digitization of the dialect layer is an important step towards preserving and enriching the national language and developing modern language technologies.

References

1. Goldin. Mashina fondi, 1986-1990.

2. Абдурахмонова, Н., & Абдувахобов, Г. (2021). O ‘quv lug ‘atini tuzishning nazariy metodologik asoslari. *Международный журнал искусство слова*, *4*(6).
3. Abdurakhmonova, N. (2021). Formal-Functional Models of The Uzbek Electron Corpus. *ANGLISTICUM. Journal of the Association-Institute for English Language and American Studies*, *10*(8), 59-66.
4. Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. In *2022 7th International Conference on Computer Science and Engineering (UBMK)* (pp. 73-75). IEEE.
5. Абдурахмонова, Н., & Абдувахобов, Г. (2021). O ‘QUV LUG ‘ATINI TUZISHNING NAZARIY METODOLOGIK ASOSLARI. *МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИСКУССТВО СЛОВА*, *4*(6).
6. Abdurakhmonova, N., Shakirovich, I. A., & O‘G‘Li, K. N. S. (2022). Morphological analyzer (morfoAnalyse) Python package for Turkic language. *Science and Education*, *3*(9), 146-156.
7. Mahmudov, M.Ə. Kompüter dilçiliyi / M.Ə. Mahmudov. – Bakı: Elm və təhsil, – 2013.– 352 s
8. Goláňová, H. Waclawičová, M. Co je v ČNK nového Ix (Zprávy z českého národního korpusu). Korpus – gramatika – axiologie, 2018 (17), pages 78–82
9. https://varieng.helsinki.fi/CoRD/corpora/Dialekts/
10. http://www.korpus.cz